Efficient Learning with Sparse Codes in Fast Hardware

Julian Göltz[(1)], Mihai A. Petrovici[(1)]

[(1)] NeuroTMA lab, Department of Physiology, University of Bern

Physical neuromorphic substrates enable efficient and fast computation; without discarding the inspiration from neuroscience, being free from certain biological restrictions gives us leverage to further optimize through innovative engineering. For example, gradient-based training allows us to maximize computational performance in spiking networks [1]. Similarly, taking inspiration from information processing in the brain, but harnessing the immense speed-up of photonic substrates, allows us to realize ultra-fast Bayesian inference through sampling [2]. Here, the biological brain works more as a toolbox rather than a paragon, from which only key characteristics are chosen for exploitation in high-efficiency applications.

DelGrad Networks of leaky integrate-and-fire neurons can be trained with exact gradient descent, even when the dynamics include trainable transmission delays [1]. Already without delays, such spiking networks work in harmony with accelerated CMOS-based hardware, enabling fast and energy-efficient classification [3]. Including delays as parameters gives another advantage: because changes in delays are different to changes in synaptic weights (Fig. 1A), they enable a qualitative jump in performance (Fig. 1B), i.e., more than could be achieved by an equivalent increase in the number of weights in the network. In addition, the on-chip delays on the neuromorphic hardware BrainScaleS-2 stabilize network dynamics, increasing performance even further (Fig. 1C).

Nanolasers Neural sampling theory builds upon the inevitable stochasticity of neural dynamics and interactions, asserting that neural activity represents samples of a world model, mediated between neurons by spikes. This essential spiking interaction is faithfully captured in nanolaser devices, with characteristic time scales on the order of nanoseconds (Fig. 1D). Simulated networks of such spiking lasers can be trained to approximate even arbitrary distributions to high fidelity (Fig. 1E). In addition to inference, like typical Boltzmann machines they can be naturally used for image generation (Fig. 1F).

Summary Both models showcase in-memory computing as well as event-driven computation, paramount characteristics of the neuromorphic approach. Furthermore, these use cases serve as examples of algorithm-hardware co-design: devising efficient and novel hardware with specific applications in mind which are in turn fine-tuned to the particular hardware. When this approach is paired with online learning *in* the physical substrate, the neuromorphic speed and energy advantages during inference can also transfer to the training process itself [4].

References

[1] Julian Göltz et al. In: *Nat. Comm.* 16 (Sept. 2025).
[2] Ivan K. Boikov et al. In: *Nat. Comm.* 17 (Dec. 2025).
[3] Julian Göltz et al. In: *Nat. Mach. Intell.* 3 (Sept. 2021).
[4] Benjamin Ellenberger et al. In: *Nat. Comm.* 17 (Dec. 2025).

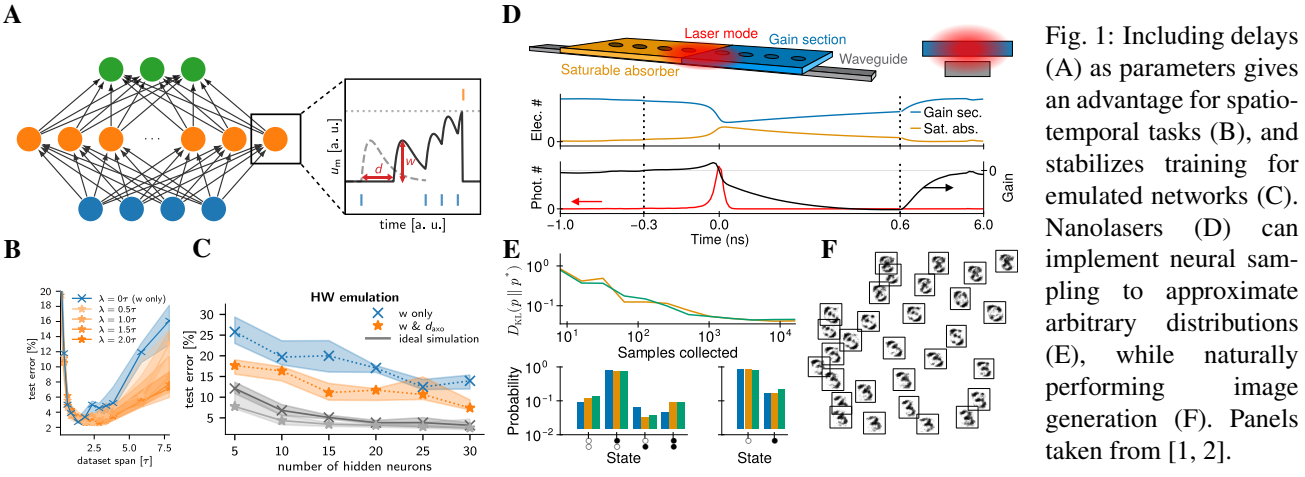

Fig. 1: Including delays (A) as parameters gives an advantage for spatio-temporal tasks (B), and stabilizes training for emulated networks (C). Nanolasers (D) can implement neural sampling to approximate arbitrary distributions (E), while naturally performing image generation (F). Panels taken from [1, 2].